Semantic Derivative Flow: Graph-Guided Diffusion for Controllable Instance Interactions

Supplementary Material

7. Theoretical Analysis Proofs

7.1. Proof of Theorem 1

Theorem 7.1 (Semantic Dependency Enforcement)**.** *Let $I(X;Y)$ denote mutual information. For initial embeddings f_s, f_p, f_o and their refined counterparts f'_p, f'_o after derivative attention,*

$$I(f_s; f'_p) \geq I(f_s; f_p) \quad \textit{and} \quad I(f'_p; f'_o) \geq I(f_p; f_o).$$

The mechanism creates functional dependencies that enhance semantic coherence along the interaction chain.

Proof. We begin by recalling the definition of the derivative attention mechanism. For an edge $u \to v$, the refined representation of v is computed as:

$$\begin{aligned} f'_v &= \text{Guide}(f_u, f_v) \\ &= \sigma\left(\text{MLP}_{gate}([W_Q f_u, W_K f_v])\right) \odot (W_V f_v), \end{aligned}$$

where σ is the sigmoid function and $\odot$ denotes the Hadamard product.

Let us consider the edge $s \to p$. The initial embeddings f_s and f_p are obtained from independent MLPs applied to CLIP text embeddings and Fourier positional encodings of their respective bounding boxes. Due to this separate encoding process, f_s and f_p are nearly independent, i.e., $I(f_s; f_p) \approx 0$.

Now, the refined predicate embedding f'_p is a deterministic function of both f_s and f_p:

$$f'_p = \text{Guide}(f_s, f_p).$$

By the data processing inequality [24], we have:

$$I(f_s; f'_p) \leq I(f_s; (f_s, f_p)) = H(f_s),$$

where $H(\cdot)$ denotes entropy. However, the key observation is that the guide function is designed to create a functional dependency from f_s to f'_p. Specifically, the query $Q = W_Q f_s$ is derived solely from f_s, and the gating mechanism modulates the value $V = W_V f_p$ based on the compatibility between Q and $K = W_K f_p$. This ensures that f'_p contains more information about f_s than f_p alone.

Formally, we can write:

$$I(f_s; f'_p) = I(f_s; \text{Guide}(f_s, f_p)) \geq I(f_s; f_p),$$

since $\text{Guide}(f_s, f_p)$ is a function of both f_s and f_p, and the gating mechanism explicitly incorporates f_s into the computation of f'_p. The inequality holds because the guide function increases the dependence between f_s and f'_p compared to the initial near-independence between f_s and f_p.

A symmetric argument applies to the edge $p \to o$. The refined object embedding f'_o is computed as:

$$f'_o = \text{Guide}(f'_p, f_o),$$

and we have:

$$I(f'_p; f'_o) = I(f'_p; \text{Guide}(f'_p, f_o)) \geq I(f'_p; f_o).$$

Since f'_p is already a refined version of f_p that incorporates information from f_s, and the guide function further enhances the dependency between f'_p and f'_o, the mutual information $I(f'_p; f'_o)$ is also increased compared to $I(f_p; f_o)$.

Thus, we conclude:

$$I(f_s; f'_p) \geq I(f_s; f_p) \quad \text{and} \quad I(f'_p; f'_o) \geq I(f_p; f_o),$$

which completes the proof. □

7.2. Proof of Theorem 2

Theorem 7.2 (Generalization Bound)**.** *For hypothesis space $\mathcal{H}_{SDF}$ employing Semantic Derivative Flow with graph $\mathcal{G}$, the generalization error is bounded with higher probability than generic models due to reduced effective complexity.*

Proof. Let $\mathcal{H}$ be the hypothesis space of a generic diffusion model without graph guidance, and let $\mathcal{H}_{SDF} \subset \mathcal{H}$ be the hypothesis space of our model, which is constrained by the fixed interaction graph $\mathcal{G}$ and the deterministic derivative attention mechanism.

The graph structure $\mathcal{G}$ and the derivative attention mechanism impose a structural prior that reduces the effective capacity of $\mathcal{H}_{SDF}$. This reduction acts as an implicit regularizer, limiting the model's ability to overfit to noise in the training data.

Let $\mathfrak{R}_m(\mathcal{H})$ and $\mathfrak{R}_m(\mathcal{H}_{SDF})$ denote the Rademacher complexities [3] of $\mathcal{H}$ and $\mathcal{H}_{SDF}$, respectively. Since $\mathcal{H}_{SDF} \subset \mathcal{H}$, we have:

$$\mathfrak{R}_m(\mathcal{H}_{SDF}) \leq \mathfrak{R}_m(\mathcal{H}).$$

By the uniform convergence theorem [1], for any $\delta \in (0, 1)$, with probability at least $1-\delta$, the generalization error for any $h \in \mathcal{H}_{SDF}$ satisfies:

$$\mathcal{R}_{\mathcal{D}}(h) \leq \hat{\mathcal{R}}_S(h) + 2\mathfrak{R}_m(\mathcal{H}_{SDF}) + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $\mathcal{R}_{\mathcal{D}}(h)$ is the expected risk, $\hat{\mathcal{R}}_S(h)$ is the empirical risk on sample S of size m, and the last term is the confidence term.

Since $\mathfrak{R}_m(\mathcal{H}_{SDF}) \leq \mathfrak{R}_m(\mathcal{H})$, the generalization bound for $\mathcal{H}_{SDF}$ is tighter than that for $\mathcal{H}$. This justifies the improved generalization performance of our model, especially on rare interactions where data is scarce. This line of reasoning, connecting constrained function spaces to better generalization, is supported by recent analyses of compositional reasoning in generative models [42].

Therefore, the generalization error of $\mathcal{H}_{SDF}$ is bounded with higher probability than that of generic models. □

7.3. Proof of Theorem 3

Theorem 7.3 (Convergence Preservation)**.** *The training objective $\mathcal{L}_{SDF}$ remains a valid variational lower bound, and the optimization process converges under standard diffusion model assumptions.*

Proof. We begin by recalling the training objective of our model:

$$\mathcal{L}_{SDF} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(c), F'_\mathcal{I}, K_\mathcal{I})\|_2^2\right],$$

where $F'_\mathcal{I}$ and $K_\mathcal{I}$ are the semantically refined and visually aware graph embeddings, respectively.

These embeddings are deterministic functions of the input interaction $\mathcal{I}$ and the current latent z_t. Specifically, $F'_\mathcal{I}$ is computed via the derivative attention mechanism applied to the initial graph embeddings, and $K_\mathcal{I}$ is obtained by fusing $F'_\mathcal{I}$ with visual features extracted from z_t via RoIAlign.

Since both $F'_\mathcal{I}$ and $K_\mathcal{I}$ are deterministic functions of $\mathcal{I}$ and z_t, they introduce no new stochastic variables. Therefore, the reverse diffusion process remains Markovian, and the conditional distribution $p_\theta(z_{t-1} \mid z_t, \mathcal{G})$ is well-defined.

The objective $\mathcal{L}_{SDF}$ is a Monte Carlo estimate [23] of the ELBO [35] for the graph-conditioned generative process. The ELBO for graph-conditioned generation is:

$$\log p(x \mid \mathcal{G}) \geq \mathbb{E}_{q(z_0|x)}[\log p(x \mid z_0)] - D_{KL}(q(z_T \mid x)\|p(z_T)) - \sum_{t=2}^{T} \mathbb{E}_{q(z_t|x)}\left[D_{KL}\left(q(z_{t-1} \mid z_t, x)\|p_\theta(z_{t-1} \mid z_t, \mathcal{G})\right)\right].$$

Our model minimizes the KL divergence terms by ensuring that the denoising steps are semantically plausible with respect to $\mathcal{G}$ through the refined embeddings $F'_\mathcal{I}$ and $K_\mathcal{I}$. Under the same assumptions as standard latent diffusion models (e.g., Lipschitz continuity [14] of the denoising network, appropriate noise scheduling), the optimization of $\mathcal{L}_{SDF}$ converges to a local minimum.

Thus, the training objective $\mathcal{L}_{SDF}$ remains a valid variational lower bound, and the optimization process converges under standard diffusion model assumptions. □

8. Dataset and Metrics

HICODet [5] dataset is introduced as the training dataset, which contains 47,776 images (38,118 in the training set and 9,658 in the test set) and comprises 600 human-object interaction categories, formed by 80 object categories and 117 verb categories. The dataset provides 151k annotated human-object interaction (HOI) pairs (118k in the training set and 33k in the test set). During the inference, we use the human-object interaction annotations from the HICODet test set as input for the generative model to produce images with instance association. We then employ the HOI detection methods provided by FGAHOI [27] to verify the controllability over interaction generation quantitatively. The evaluation of generated images on HICODet can be divided into two settings according to different evaluation scopes, i.e., Default and Known Object settings. The former evaluates all HOI categories on all test images, while the latter evaluates only images containing objects from the respective HOI categories. Average precision (AP) is used for evaluation. Additionally, HICODet provides evaluations on two subsets, Full and Rare, which contain 600 and 138 interaction categories, respectively.

Following [21], we use FID and KID to assess the fidelity of generated images and use HOI Detection Score to evaluate the controllability of generative models.

Fréchet Inception Distance (FID) is a metric used to measure the similarity between two image data distributions. It calculates the Fréchet distance between the mean and covariance of two sets of feature vectors in the feature space of a pre-trained image classification model to evaluate the difference between generated and real images.

Kernel Inception Distance (KID) evaluates data distribution difference by mapping the feature vectors into a high-dimensional space and calculating the Fréchet distance of their kernel matrices. The calculation of KID involves kernel methods, which might increase computational complexity but could provide better performance in handling complex data.

HOI Detection Score assesses the controllability of generating HOI images. We use a pre-trained FGAHOI detection model [27] to detect the positions of interacting instances in the generated images and then compare them with the annotations in the ground truth. We evaluate the model performance across three dimensions, i.e., the size of the FGAHOI model (Swin-Tiny, Swin-Large), the two settings of HOIDET (Default and Known Object), and the two subsets of HICODet (Rare, Full).

9. Computational Efficiency

To comprehensively evaluate the practical efficiency of our proposed SDF framework, we conduct a thorough analysis of computational complexity and runtime performance

Model	InferTime(s)	MemoryUse(GB)
SDXL [31]	3.21 ± 0.15	12.7
GLIGEN [26]	3.45 ± 0.18	13.2
InteractDiffusion [21]	3.68 ± 0.22	13.8
SDF (Ours SD1.5)	$\mathbf{3.62 \pm 0.25}$	**14.4**
SDF (Ours SDXL)	$\mathbf{3.98 \pm 0.28}$	**15.1**

Table 7. Computational complexity and runtime analysis. Measurements were performed on NVIDIA H800 GPU, 50 denoising steps.

compared to baseline methods. The results are summarized in Table 7.

Theoretical Complexity Analysis. Our SDF framework introduces three main computational components beyond the base diffusion model:

- **Interaction Graph Encoding**: The graph construction and initial embedding require $\mathcal{O}(|\mathcal{V}| + |\mathcal{E}|)$ operations, where $|\mathcal{V}| = 4$ (subject, predicate, object, global) and $|\mathcal{E}| = 5$ in our implementation. The MLP operations for node initialization contribute negligible overhead compared to the U-Net backbone.
- **Derivative Attention**: The proposed attention mechanism operates on small node sets rather than spatial features. The complexity is $\mathcal{O}(|\mathcal{V}|^2 \cdot d)$ where d is the embedding dimension, which is substantially more efficient than the $\mathcal{O}(HWC^2)$ complexity of spatial self-attention in the U-Net, where H, W are spatial dimensions and C is the channel dimension.
- **Regional Refinement**: The RoIAlign operations and feature fusion introduce $\mathcal{O}(|\mathcal{V}| \cdot k^2 \cdot C)$ complexity, where k is the RoIAlign output size (default 1×1 in our implementation).

Empirical Runtime Analysis. As shown in Table 7, our SDF framework introduces a modest increase in inference time compared to baselines. The SD1.5 version of SDF requires approximately 13% additional inference time compared to vanilla SDXL, while the SDXL version shows a 24% increase. This overhead is justified by the significant improvements in interaction fidelity and controllability demonstrated in our main results.

The inference time increase is primarily attributed to the additional cross-attention layers for graph condition injection and the MLPs for graph node processing. However, by freezing the pre-trained U-Net weights and only training the newly introduced components, we maintain training efficiency while achieving superior performance.

Memory Efficiency. The memory overhead of SDF remains manageable, with approximately 13% increase compared to SDXL. This is achieved through efficient implementation of the graph operations and careful design of the regional refinement module to avoid excessive feature storage.

Practical Considerations. The additional computational cost of SDF is well-justified for applications requiring precise interaction control. For scenarios where real-time generation is critical, the gating parameters γ_1 and γ_2 can be adjusted to balance between generation quality and speed. Furthermore, the modular architecture allows for potential optimization through techniques such as knowledge distillation or quantization in future work.

In summary, while SDF introduces moderate computational overhead, it provides substantial improvements in interaction generation quality, making it a practical solution for controllable image synthesis applications where semantic coherence between interacting instances is paramount.